Session 3

Describing Distributions

Key Terms for This Session

Previously Introduced

- cumulative frequency
- discrete data
- frequency
- frequency bar graph
- interval
- line plot
- median
- mode
- relative frequency
- relative frequency bar graph
- variation

New in This Session

- continuous variable
- grouped frequency table
- histogram
- relative frequency histogram
- stem and leaf plot

Introduction

In the previous session, you saw how different ways of representing data—such as line plots, bar graphs, frequency and relative frequency tables, and cumulative frequency and relative cumulative frequency tables—allowed you to provide better answers to statistical questions.

We also examined how to answer statistical questions when there is variation in data. One idea was to express your answer as an interval, such as the interval in which all of the data are located or an interval with a concentration of data. Another method was to use a "typical" value to represent all the values in your data set, such as the mode or the median.

In this session, you will investigate some approaches to grouping data in graphs and tables, and you will examine different types of statistical answers to questions based on these grouped representations.

Learning Objectives

In this session, you will learn how to do the following:

- Organize data in a stem and leaf plot
- Group the data from a stem and leaf plot
- Complete a frequency and relative frequency table for your grouped data
- Create a frequency and relative frequency histogram for your grouped data
- Complete a cumulative frequency and relative cumulative frequency table for your grouped data

Part A: Organizing Data in a Stem and Leaf Plot (55 min.)

How Long Is a Minute?

Let's begin with a problem you saw in Session 1, the "How Long Is a Minute?" problem. You will use the data from this problem to create a stem and leaf plot, a useful device for organizing certain types of data. **[See Note 1]**

As always, we begin with Step 1 of our four-step problem-solving process.

Ask a question:

How good is your sense of time? Without a timing device, how well can you judge the actual length of a minute? Are some people better at judging elapsed time than others?

Collect appropriate data:

Twenty-six people tried this activity. At the end of what each person judged to be a minute, the actual time that had elapsed was recorded to the nearest second. The responses (in seconds) were as follows:

63	67	79	75
57	72	52	89
39	59	55	68
66	86	70	52
60	64	42	54
56	82	57	65
59	33		

Note that this is *quantitative* data, since the responses are numerical values. Time, however, is not obtained by counting as you did when you determined the number of raisins in a box. Time can be measured on a number line, and any point on the line is a possible point in time. The recorded times above were rounded off to the nearest second. But any positive real number is a possible measurement for time. In this way, time is a continuous variable, and data collected on this type of variable are called continuous data. This is in contrast to a discrete numerical variable (like the raisins), which is often obtained by counting and usually assumes only whole numbers as values.

Another example of a continuous variable is height, measured in centimeters. A person's height can be any positive number, even though the data are typically rounded off to the nearest centimeter.

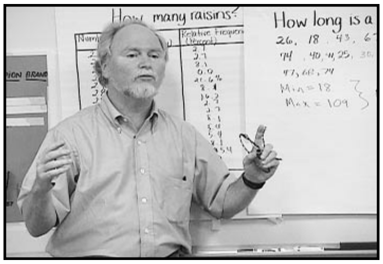
Organizing continuous data, or discrete data with a great deal of variation, often requires that values be grouped.

Analyze the data:

Problem A1. Can you think of why a line plot might not be a useful way to illustrate this data set?

Note 1. Data are provided for this activity; however, if you collected your own data in Session 1, you can use that data instead. Keep in mind the nature of this data—it is *quantitative*. Any positive number could be a measurement of time, although you can choose to round your data to the nearest second, as we've done with the provided data.

Part A, cont'd.



Video Segment (approximate times: 2:24-3:05): You can find this segment on the session video approximately 2 minutes and 24 seconds after the Annenberg/CPB logo. Use the video image to locate where to begin viewing.

In this video segment, participants discuss why a line plot would not be a useful way to display the results of their statistical inquiry. They then discuss how grouping the data could provide them with more information. Watch this segment after completing Problem A1.

What would a more useful graphical representation of the data look like?

Making a Stem and Leaf Plot

Since we've decided that a line plot may not be a useful graph for investigating variation, we must come up with a new representation based on groups of data. One such representation is called a stem and leaf plot. **[See Note 2]**

Let's look at our data again:

63	67	79	75
57	72	52	89
39	59	55	68
66	86	70	52
60	64	42	54
56	82	57	65
59	33		

Notice that these data consist of two-digit numbers. The smallest data value is 33, the largest is 89, and there are data values in the 30s, 40s, 50s, 60s, 70s, and 80s. It would be reasonable to group the data by tens, and a stem and leaf plot can help us do this.

The first step is to organize the data in groups of 10. Each *stem* of a stem and leaf plot is determined from the left-most digit(s) of each number—in this case, this is the tens digit. For example, the stem of the first data value (63) is 6, and the stem of the data value below it (57) is 5.

To construct the stem and leaf plot, start by listing all possible stems within the range of the data:

```
Stem
 3 |
 4 |
 5 |
 6 |
 7 |
 8 |
```

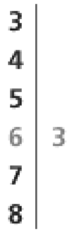
The *leaf* of each data value in a stem and leaf plot is determined from the rightmost digit of each number—in this case, this is the units digit. For example, the leaf of the first data value (63) is 3, and the leaf of the data value below it (57) is 7.

Note 2. As you create the stem and leaf plot, notice that this organization of the data is based on grouping by tens. Think about how patterns that appear in the stem and leaf plot would not appear in a line plot.

Part A, cont'd.

To construct a stem and leaf plot, go through the data list one value at a time, and record the leaf of each number beside the proper stem. For example, the first data value, 63, has stem 6 and leaf 3:

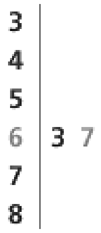
63	67	79	75
57	72	52	89
39	59	55	68
66	86	70	52
60	64	42	54
56	82	57	65
59	33		



3	
4	
5	
6	3
7	
8	

If we move across the top row, the second data value (67) has stem 6 and leaf 7. Put this value next to the first leaf (3) already listed on stem 6. (The leaf entries do not have to be ordered at this time.)

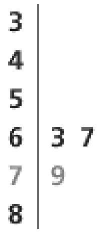
63	**67**	79	75
57	72	52	89
39	59	55	68
66	86	70	52
60	64	42	54
56	82	57	65
59	33		



3	
4	
5	
6	3 7
7	
8	

The next data value (79) has stem 7 and leaf 9:

63	67	**79**	75
57	72	52	89
39	59	55	68
66	86	70	52
60	64	42	54
56	82	57	65
59	33		



3	
4	
5	
6	3 7
7	9
8	

Problem A2. Continue constructing the stem and leaf plot.

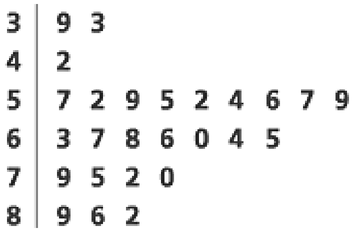
Try It Online! **www.learner.org**

This problem can be explored online as an Interactive Activity. Go to the *Data Analysis, Statistics, and Probability* Web site at www.learner.org/learningmath and find Session 3, Part A, Problem A2.

Part A, cont’d.

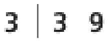
Ordering a Stem and Leaf Plot

If we work through the entire data set, our initial stem and leaf plot looks like this:



```
3 | 9 3
4 | 2
5 | 7 2 9 5 2 4 6 7 9
6 | 3 7 8 6 0 4 5
7 | 9 5 2 0
8 | 9 6 2
```

The final step in creating a stem and leaf plot is to order the data within each stem. For example, the first stem, 3, has two leaves, 9 and 3. To order this list of leaves, arrange them in order from 0 to 9. The first row becomes:



```
3 | 3 9
```

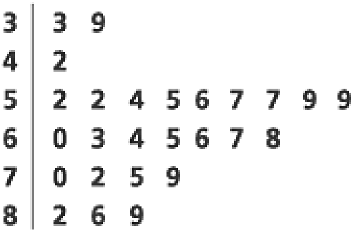
Problem A3. Create an ordered stem and leaf plot of the data.

Try It Online! **www.learner.org**

This problem can be explored online as an Interactive Activity. Go to the *Data Analysis, Statistics, and Probability* Web site at www.learner.org/learningmath and find Session 3, Part A, Problem A3.

Interpreting the Stem and Leaf Plot

The ordered stem and leaf plot from Problem A3 looks like this:



```
3 | 3 9
4 | 2
5 | 2 2 4 5 6 7 7 9 9
6 | 0 3 4 5 6 7 8
7 | 0 2 5 9
8 | 2 6 9
```

A stem and leaf plot shows us potential patterns in the responses that may not be apparent in the original listing of the data. For example, we can see that a large number of data values are in the 50s and 60s. Ordering a stem and leaf plot offers another way to represent the answers to our question, “How well do people judge when a minute has elapsed?”

Part A, cont'd.

Problem A4. Based on the ordered stem and leaf plot:

a. How many of the estimates are between 33 and 89 seconds (inclusive)?

b. How many of the estimates are between 52 and 68 seconds (inclusive)?

Problem A5.

Since the goal is to estimate when a minute has elapsed, it makes sense to consider how close the estimates are to the correct response, which is 60 seconds.

a. How many people's estimates were more than five seconds away from one minute? That is, how many of the responses were less than 55 seconds or greater than 65 seconds?

b. How many estimates were within five seconds of one minute?

c. How many estimates were more than 10 seconds away from one minute?

d. How many estimates were within 10 seconds of one minute?

Problem A6.

a. Determine the mean of this data set. How does the mean compare to the correct response of 60 seconds?

b. How many people's estimates were more than five seconds away from the mean?

c. How many people's estimates were more than 10 seconds away from the mean?

d. Why is it not useful to calculate the mode for this data set? **[See Tip A6, page 86]**

Asking and answering questions like the ones in Problems A5 and A6 can help us learn more about the variation present in a data set. They are important questions to consider as we interpret our data.

Grouping by Fives

The stem and leaf plot for the 26 estimates of elapsed time illustrates a grouping of the data by tens; for example, the first stem contains all values from 30 to 39. A stem and leaf plot, however, does not have to group the data by tens—we could have grouped by fives, for instance. If we were grouping by fives, we would consider all possible numbers in the 50s, for instance, and then put them in two groups, the High and the Low:

Low Group	**High Group**
50 51 52 53 54	55 56 57 58 59

When forming the stems for a grouping by fives, we consider the second digit of each number as well as the first:

- Numbers in the Low Group end with a second digit of 0, 1, 2, 3, or 4.
- Numbers in the High Group end with a second digit of 5, 6, 7, 8, or 9.

Note that the two groups each have the same number of possible second digits (5). When creating a stem and leaf plot, all stems should have the same number of possible leaves.

Part A, cont'd.

To classify our 26 responses in this way, we set up our stem and leaf plot with stems corresponding to the Highs (H) and Lows (L) and then group the responses accordingly. For example, the first data value (63) is on stem 6L since its leaf, 3, is in the Low Group:

63	67	79	75
57	72	52	89
39	59	55	68
66	86	70	52
60	64	42	54
56	82	57	65
59	33		

3L	
3H	
4L	
4H	
5L	
5H	
6L	3
6H	
7L	
7H	
8L	
8H	
9L	
9H	

Try It Online! **www.learner.org**

Try It Online! This problem can be explored online as an Interactive Activity. Go to the *Data Analysis, Statistics, and Probability* Web site at www.learner.org/learningmath and find Session 3, Part A, Problem A7.

Problem A7. Convert the stem and leaf plot grouped by tens into a stem and leaf plot grouped by fives.

Ordering Low and High

Before ordering, the stem and leaf plot based on a grouping by fives looks like this:

3L	3
3H	**9**
4L	2
4H	
5L	2 2 4
5H	**7 9 5 6 7 9**
6L	3 0 4
6H	**7 8 6 5**
7L	2 0
7H	**9 5**
8L	2
8H	**9 6**

After ordering, the stem and leaf plot based on a grouping by fives looks like this:

3L	3
3H	**9**
4L	2
4H	
5L	2 2 4
5H	**5 6 7 7 9 9**
6L	0 3 4
6H	**5 6 7** 8
7L	0 2
7H	**5 9**
8L	2
8H	**6 9**

Part A, cont'd.

When the stem and leaf plots for the two different groupings are placed next to each other, you can see the connections between the two as well as some different patterns in the variation. Each stem of the first stem and leaf plot corresponds to two stems in the second: one that represents the lower five digits in the leaves and one that represents the upper five. Increasing the number of stems (e.g., five units per stem rather than 10) allows you to see smaller degrees of variation between stems, but each stem will have fewer leaves. You must find the best compromise between stems that are too wide to differentiate between data and stems that are too narrow to see trends in the overall data.

3	3 **9**
4	2
5	2 2 4 **5 6 7 7 9 9**
6	0 3 4 **5 6 7 8**
7	0 2 **5 9**
8	2 **6 9**

3L	3
3H	**9**
4L	2
4H	
5L	2 2 4
5H	**5 6 7 7 9 9**
6L	0 3 4
6H	**5 6 7** 8
7L	0 2
7H	**5 9**
8L	2
8H	**6 9**

You can choose different-sized groupings for the stem and leaf plots for different data sets (e.g., one can be grouped by fives and one can be grouped by 100s, if those are the groupings that will work best). In general, try to use no fewer than five and no more than 15 stems when constructing a stem and leaf plot.

Problem A8. Based on the stem and leaf plot grouped by fives, give two descriptive statements that provide an answer to the question "How well do people judge when a minute has elapsed?" Your answers should take into account the variation in the data.

Problem A9.

a. Think of a situation in which it would be useful to create a stem and leaf plot that would be grouped by a number larger than 10.

b. Think of a situation in which a stem and leaf plot would be impractical or would not be an effective way to present your data.

Part B: Histograms (30 min.)

Constructing a Histogram

Like the line plot we explored in Session 2, the stem and leaf plot is a useful device for illustrating variation in data for small data sets (up to 100 values). For larger data sets, though, the stem and leaf plot is not a practical way to organize data. Instead, you might want to use a histogram. **[See Note 3]**

Let's start with the stem and leaf plot for a new data set: 52 estimates collected in answer to the question "How long is a minute?":

79	67	72	75	64	82
55	56	58	66	60	59
63	75	66	57	72	
57	62	57	67	53	
67	59	61	60	57	
61	60	53	30	50	
42	39	68	89	67	
65	86	39	54	93	
52	55	72	56	65	
89	33	52	60	70	

```
3 | 0 3 9 9
4 | 2
5 | 0 2 2 3 3 4 5 5 6 6 7 7 7 7 8 9 9
6 | 0 0 0 0 1 1 2 3 4 5 5 6 6 7 7 7 7 8
7 | 0 2 2 2 5 5 9
8 | 2 6 9 9
9 | 3
```

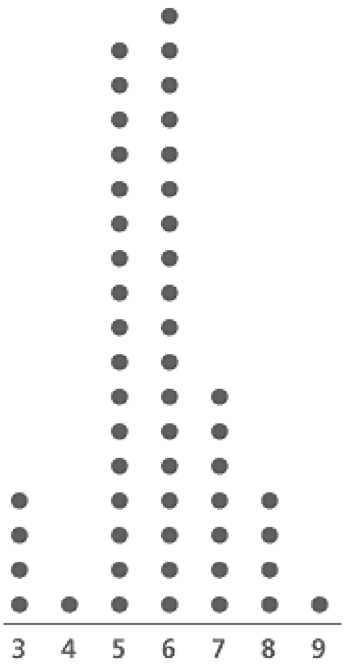
If the stem and leaf plot is rotated 90° counterclockwise, it looks like this:

```
    8 7
    7 7
    7 6
    7 6
    7 5
    6 5
    6 4
    5 3
    5 2 9
    4 1 5
    3 1 5
9   3 0 2 9
9   2 0 2 9
3   2 0 2 6
0 2 0 0 0 2 3
-------------
3 4 5 6 7 8 9
```

Note 3. You will take an evolutionary approach to developing the histogram. The objective of this activity is for you to see the relationship between the line plot, the stem and leaf plot, and the histogram. The line plot shows the frequencies of the rows, but not the actual data values. The stem and leaf plot contains more detailed information than the histogram in that all of the data values are shown. And finally, the relative frequency histogram shows the relative sizes of the frequencies for each interval, although it does not explicitly show those frequencies.

Part B, cont'd.

To create a histogram for this data, first replace each "leaf" (second digit) with a dot:

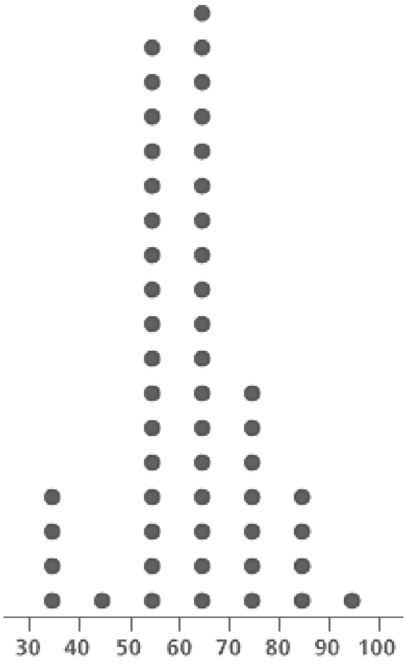


While a histogram is similar to a line plot, in fact, there are differences in the values across the horizontal axis. In a line plot, these numbers represent a single data value. In the plot at left, the numbers across the bottom indicate the stems in the original stem and leaf plot. Each number represents an entire interval of values.

For instance, the "3" denotes the stem for all values in the 30s—that is, the interval (range) of values from 30 up to (but not including) 40. For the purposes of a histogram, it is useful to label this interval "30 to less than 40" (30 to < 40) to remind us that 30 is included but 40 is not.

The "4" denotes the stem for all values in the 40s—that is, the interval (range) of values from 40 up to (but not including) 50. Again, it is helpful to label this interval "40 to < 50" to remind us that 40 is included but 50 is not.

If we re-label the horizontal axis to show these intervals (groups) of data, the graph below is produced. Again, this graph is similar to a line plot except that the horizontal axis indicates intervals of data values instead of individual data values:



A grouped frequency table can be determined from this display in the following manner:

- There are four dots over the first group in the interval 30 to < 40. This group has frequency 4.
- There is one dot over the second group in the interval 40 to < 50. This group has frequency 1.

Continue this process for the other groups. Finally:

- There is one dot over the final group in the interval 90 to < 100. This group has frequency 1.

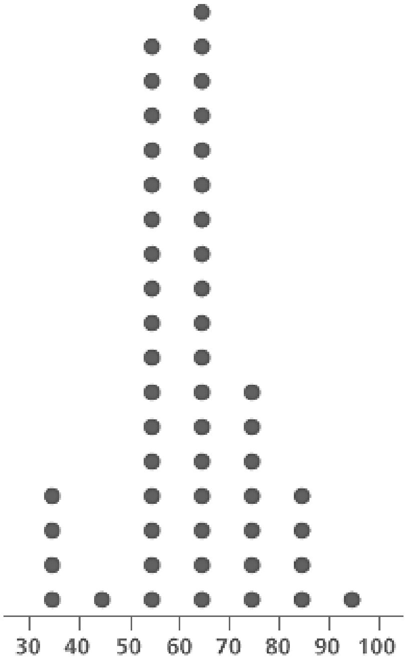
Part B, cont'd.

This process produces the following grouped frequency table:

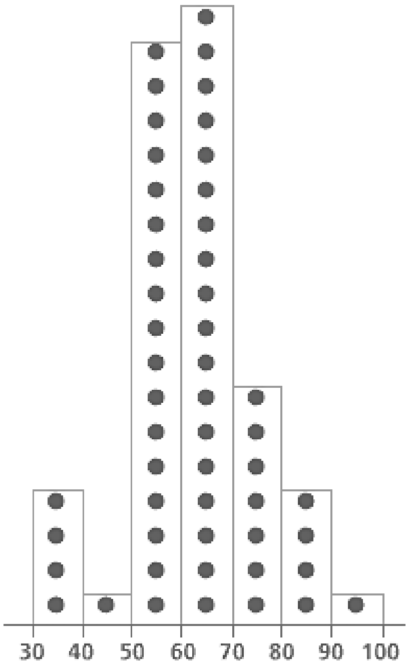
Interval	Frequency for Interval
30 to < 40	4
40 to < 50	1
50 to < 60	17
60 to < 70	18
70 to < 80	7
80 to < 90	4
90 to < 100	1

Remember that this table describes *ranges* of data values rather than specific data values. For instance, we can see that there are seven responses in the interval 70 to < 80, but we have no idea what the actual values are for those responses.

You are now ready to construct a histogram from the dot plot you have created:



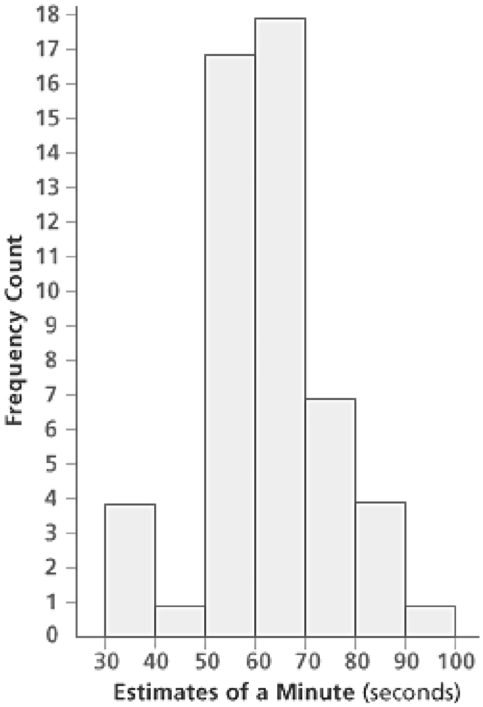
- Draw a rectangle over each value on the horizontal axis with a height corresponding to the frequency of that value:



Note that the frequency of each value on the horizontal axis is still indicated by the number of dots within each rectangle.

Part B, cont'd.

- Remove the dots, shade the rectangles, and add a vertical scale to indicate the frequency of each interval on the horizontal scale:



Try It Online! **www.learner.org**

Try It Online! The ideas presented here can be explored online as an Interactive Activity. Go to the *Data Analysis, Statistics, and Probability* Web site at www.learner.org/learningmath and find Session 3, Part B.

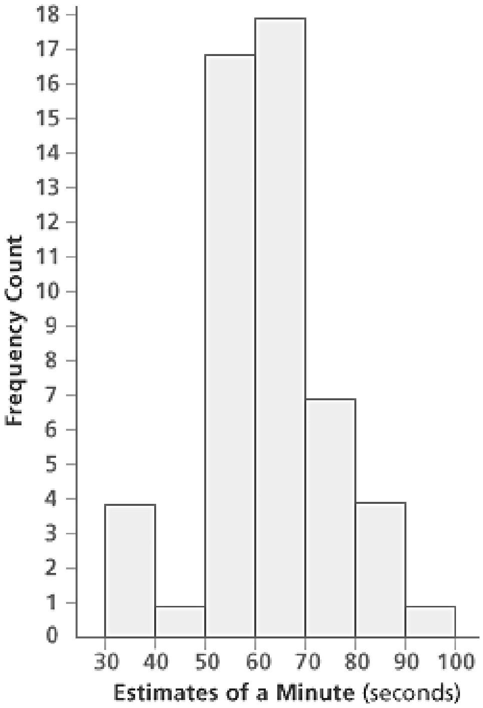
You have just created a frequency histogram!

Problem B1. What advantages does a histogram have over a stem and leaf plot? What are the disadvantages of a histogram?

As with the sizes of each group in a stem and leaf plot, you can change the size of the intervals in your histogram depending on the situation. Your ongoing goal is to present your data in the most effective way possible in order to extract meaning from the data, and the histogram can be a useful tool for doing so.

Part B, cont'd.

Interpreting a Histogram



Interval	Frequency for Interval
30 to < 40	4
40 to < 50	1
50 to < 60	17
60 to < 70	18
70 to < 80	7
80 to < 90	4
90 to < 100	1

The histogram and grouped frequency table you just created offer different ways to present your data (the time estimates) and provide different ways to answer our original question,"How well do people judge when a minute has elapsed?"

Problem B2. Using only the histogram and grouped frequency table, give two descriptive statements that provide an answer to this question. (Since the goal is to estimate when a minute has elapsed, it would make sense to again consider how close the estimates are to the correct response—60 seconds.)

Problem B3.

a. According to the histogram and grouped frequency table, how many people's estimates were outside the interval from 50 to less than 70 seconds? That is, how many estimates were less than 50 seconds or 70 seconds or more?

b. How many estimates were within the interval from 50 to less than 70 seconds?

c. How many estimates were outside the interval from 40 to less than 80 seconds?

d. In Problem A5, only nine people's estimates were more than 10 seconds away from one minute. Does your answer to question (a) of this problem imply that the people in this group were not as good at estimating a minute's time? If so, why? If not, how could you make a fairer comparison between the two sets? **[See Tip B3, page 86]**

Part C: Relative and Cumulative Frequencies (30 min.)

Relative Frequencies

The frequency histogram and grouped frequency table for the 52 time estimates contain similar information to the stem and leaf plot, but they don't indicate each person's actual estimates. The height of each bar in the histogram indicates the frequency of the corresponding interval of estimates on the horizontal axis. **[See Note 4]**

As with the stem and leaf plot, the frequency histogram can be an awkward graph for large data sets, since the vertical axis corresponds to the frequency of each interval of values. For large data sets, some intervals may have many values and a high frequency. Consequently, the vertical axis would have to be scaled according to the largest frequency.

An alternative is to use relative frequencies to describe how many values are in each interval relative to the total number of values. For most purposes, relative frequencies are more useful than absolute frequencies; for example, the statement "17 of the 52 estimates are in the interval 50 to < 60" is more useful than the statement "17 estimates are in the interval 50 to < 60."

The relative frequency for the interval 50 to < 60 is 17/52, which you can also write in decimal form as .327 (rounded to three digits). Multiplying by 100 gives you the percentage, 32.7%. This means that 32.7% of the estimates are in the interval 50 to < 60.

Here is what you get for the rest of the data:

Interval	Frequency	Relative Frequency		
		Fraction	Decimal	Percentage
30 to < 40	4	4/52	.077	7.7
40 to < 50	1	1/52	.019	1.9
50 to < 60	17	17/52	.327	32.7
60 to < 70	18	18/52	.346	34.6
70 to < 80	7	7/52	.135	13.5
80 to < 90	4	4/52	.077	7.7
90 to < 100	1	1/52	.019	1.9

Notice that the relative frequencies expressed as fractions and decimals add up to 1 and that the percentages add up to 100%.

Note 4. Cumulative frequencies and relative cumulative frequencies are introduced in Part C, using all three representations—the line plot, the stem and leaf plot, and the histogram. Again, seeing this idea in the different representations not only reinforces what you learn about the data set, but also emphasizes the relationships between the different representations.

Part C, cont'd.

Problem C1. Use only the relative frequencies from the table to answer the questions below. Give your answers as percentages, to the nearest 10th of a percent, or explain why the answer cannot be found from the table.

a. What percentage of the responses are in the 70s and below?

b. What percentage of the responses are 80 or higher?

c. What percentage of the responses are in the 50s and below?

d. What percentage of the responses are 60 or higher?

e. What percentage of the responses are less than 100?

f. What percentage of the responses are at least 40 but below 70?

g. What percentage of the responses are 65 or greater?

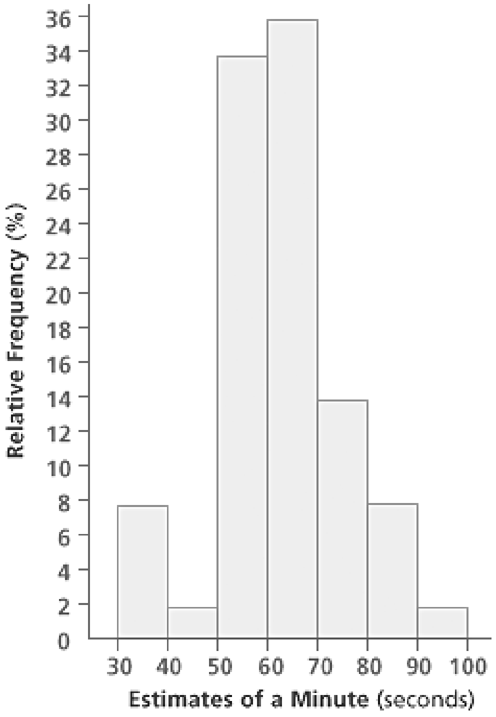
h. What percentage of the responses are less than 35?

i. What percentage of the responses are equal to 60? **[See Tip C1, page 86]**

Take It Further

Problem C2. For questions (g) and (h) in Problem C1, use the table to come up with an estimated percentage. **[See Tip C2, page 86]**

The relative frequency histogram looks similar to the frequency histogram; the only differences are that the labels along the vertical axis represent percentages, and the height of each bar now represents the relative frequency expressed as a percentage for the corresponding interval of values. Here is the relative frequency histogram for the 52 estimates of a minute:



Part C, cont'd.

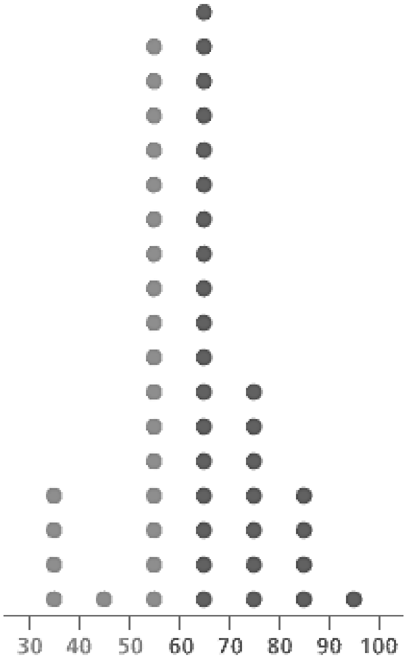
Cumulative Frequencies

As we did in Session 2, we can define cumulative frequencies based on intervals of data. For example, the number of responses that are less than 60 is the cumulative frequency of 60. (Note that "below 60" means "in the 50s and below.")

If you begin with the stem and leaf plot for the 52 estimates of a minute, there are 22 values below 60:

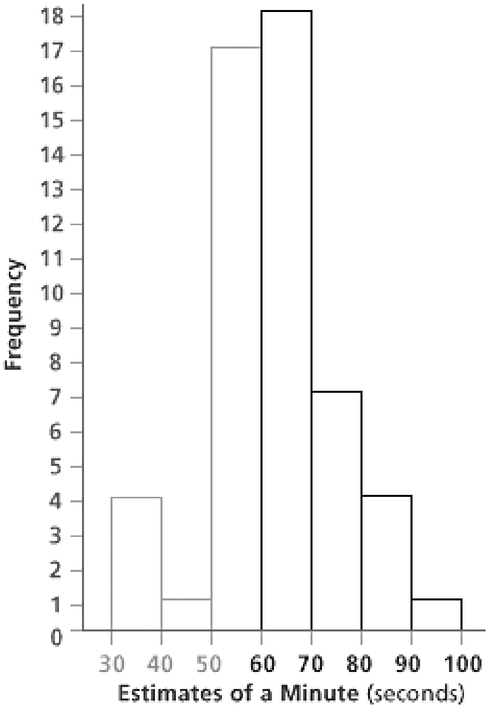
```
      8
    9 7
    9 7
    8 7
    7 7
    7 6
    7 6
    7 5
    6 5
    6 4
    5 3
    5 2 9
    4 1 5
    3 1 5
9   3 0 2 9
9   2 0 2 9
3   2 0 2 6
0 2 0 0 0 2 3
-------------
3 4 5 6 7 8 9
```

The corresponding 22 dots are shown in the dot version of this stem and leaf plot:



Part C, cont'd.

And finally, the corresponding bars in the frequency histogram are indicated below:



These three representations tell us that there are 22 estimates below 60.

Problem C3. Complete this cumulative frequency table with the information you collected from the histograms above:

Interval	Frequency	Cumulative Frequency
30 to < 40	4	
40 to < 50	1	
50 to < 60	17	22
60 to < 70	18	
70 to < 80	7	
80 to < 90	4	
90 to < 100	1	

Part C, cont'd.

Problem C4. Use only the cumulative frequencies from the table to answer the questions below. As with Problem C1, first determine whether a question can be answered using only this table.

a. How many responses are in the 70s and below?

b. How many responses are 80 or higher?

c. How many responses are in the 50s and below?

d. How many responses are 60 or higher?

e. How many responses are less than 100?

f. How many responses are at least 40 but below 70?

g. How many responses are 65 or greater?

h. How many responses are less than 35?

i. How many responses are equal to 60?

Relative Cumulative Frequencies

You can convert cumulative frequencies to relative cumulative frequencies by dividing each cumulative frequency by the total number of data values. For the 52 estimates of a minute, here are the relative cumulative frequencies:

Interval	Cumulative Frequency	Relative Cumulative Frequency		
		Fraction	Decimal	Percentage
30 to < 40	4	4/52	.077	7.7
40 to < 50	5	5/52	.096	9.6
50 to < 60	22	22/52	.423	42.3
60 to < 70	40	40/52	.769	76.9
70 to < 80	47	47/52	.904	90.4
80 to < 90	51	51/52	.981	98.1
90 to < 100	52	52/52	1.000	100.0

Problem C5. Use only the relative cumulative frequencies from this table to answer the questions below. This time, give your answers in percentages, to the nearest 10th of a percent.

a. What percentage of the responses are in the 70s and below?

b. What percentage of the responses are 80 or higher?

c. What percentage of the responses are in the 50s and below?

d. What percentage of the responses are 60 or higher?

e. What percentage of the responses are less than 100?

f. What percentage of the responses are at least 40 but below 70?

Part D: Ordering Hats (35 min.)

Understanding the Question

In Parts A-C of this session, you learned several ways to organize numerical data by forming groups. Grouping is especially useful for wide-ranging data or data measured on the number line. **[See Note 5]**

In the following activity, you will apply several of the methods you have learned for grouping data to solve a problem about how many hats to order.

Hats are made in a variety of styles and sizes. A merchant must decide what styles to keep in stock and how many of each size to order. At our theoretical hat shop, a unisex "Standard Fit" hat is available in the following sizes:

Standard Size	Fits Head Circumference (mm)
S1	520 to < 530
S2	530 to < 540
S3	540 to < 550
S4	550 to < 560
S5	560 to < 570
S6	570 to < 580
S7	580 to < 590
S8	590 to < 600
S9	600 to < 610
S10	610 to < 620

Are certain hat sizes more common than others? If not, then an equal number of each hat size can be ordered. But if certain sizes are more common, the merchant needs to order larger quantities of the more common sizes.

Write and Reflect

Problem D1. Before you begin, make an initial guess about whether you expect all hat sizes to be equally common. Explain your answer. If you think some hat sizes will be more common than others, which hat size would you expect to be the most common, and why?

Video Segment (approximate times: 7:46-8:45): You can find this segment on the session video approximately 7 minutes and 46 seconds after the Annenberg/CPB logo. Use the video image to locate where to begin viewing.

In this video segment, the hat-ordering problem is introduced, and participants describe their initial expectations for a hat-size distribution. Watch this segment after completing Problem D1.

Note 5. If you are working in a group, use your own data for the Ordering Hats activity. Each person should measure the head circumferences of several adults ahead of time, then bring their data to class. The group should have a total of 50 to 60 head circumferences for their data set. Also, consider having each person measure an equal number of men's and women's head circumferences. As an extension, you can look at the data separately for each sex.

Fathom Dynamic Statistics™ Software, used by the onscreen participants, is helpful in creating graphical representations of data. You can use Fathom Software to complete Problems D3-D13, as well as Homework Problems H1-H6. For more information, go to the Key Curriculum Press Web site at http://www.keypress.com/fathom/.

Part D, cont'd.

How could a hat merchant determine which hat sizes are most common?

Hat size is clearly determined by head size. Several possible measurements of the human head might be used to describe head size. Mail-order catalogs ask you to measure your head circumference and then determine your hat size from a table of circumferences they provide. To find your head circumference, measure the largest part of your head by placing the measuring tape just above your eyebrows.

Problem D2. What size Standard Fit hat would you wear?

Data Analysis Using a Stem and Leaf Plot

Let's plan an order for 1,000 Standard Fit hats.

Ask a question:

How large are people's heads? Are some head sizes more common than others?

For an order of 1,000 unisex Standard Fit hats, how many of each size should you order?

Collect appropriate data:

We used a metric tape measure to measure the head circumferences of 55 people to the nearest millimeter:

615	542	550	580	590	540	566	608	555	556	580
580	562	564	578	580	600	565	555	568	590	548
577	579	555	569	603	560	550	587	556	584	590
603	554	560	569	532	570	600	590	557	607	560
559	570	534	520	560	554	610	600	600	570	560

Analyze the data:

Problem D3. Create a stem and leaf plot for these data using stems that correspond to Standard Fit hat sizes. Keep in mind that these are three-digit numbers and that, for these data, the stems will be based on the left two digits of the values.

Interpret the results:

Problem D4. Based on the stem and leaf plot, what are some things you can say about the way head sizes are distributed? Give two descriptive statements to answer the question "How large are people's heads?"

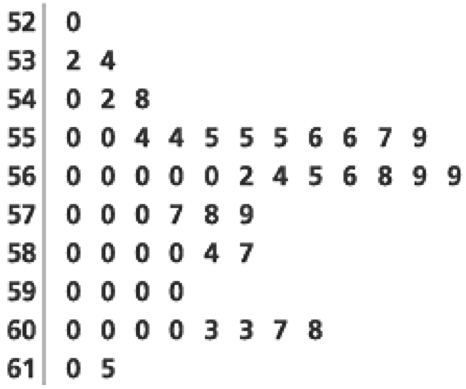
Problem D5. Based on the stem and leaf plot, do some head sizes appear to be more common than others? Which head sizes are most common? Least common?

Problem D6. What would happen if you ordered an equal number of each size of Standard Fit hats?

Part D, cont'd.

Using a Histogram To Analyze the Hat-Size Data

The stem and leaf plot for this data set looks like this:



```
52 | 0
53 | 2 4
54 | 0 2 8
55 | 0 0 4 4 5 5 5 6 6 7 9
56 | 0 0 0 0 0 2 4 5 6 8 9 9
57 | 0 0 0 7 8 9
58 | 0 0 0 0 4 7
59 | 0 0 0 0
60 | 0 0 0 0 3 3 7 8
61 | 0 5
```

Analyze the data:

Problem D7. Use the stem and leaf plot to determine the following:

a. The grouped frequency and relative frequency tables for the head-circumference data

b. The frequency histogram for the head-circumference data

Problem D8. What information in the data is "lost" when the distribution is represented by a grouped frequency table and histogram instead of a stem and leaf plot?

Interpret the results:

Problem D9. What can you say about the way head sizes are distributed? Based on the grouped relative frequency table and the histogram, give two descriptive statements to answer the question "How large are people's heads?"

Problem D10. Based on the grouped frequency table and the histogram, do some head sizes appear to be more common than others? Which head sizes are most common? Least common?

Problem D11. Based on these data, plan an order for 1,000 Standard Fit hats. **[See Tip D11, page 86]**

Standard Size	Fits Head Circumference (mm)	Number To Order
S1	520 to < 530	
S2	530 to < 540	
S3	540 to < 550	
S4	550 to < 560	
S5	560 to < 570	
S6	570 to < 580	
S7	580 to < 590	
S8	590 to < 600	
S9	600 to < 610	
S10	610 to < 620	

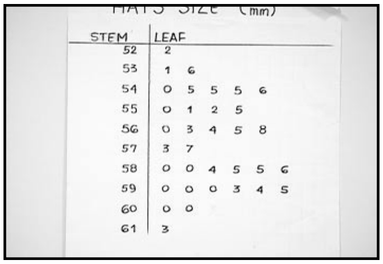
Part D, cont'd.

Problem D12. Using frequency computations, the total number of hats might not be exactly 1,000.

a. Why did this happen?

b. To complete the order of 1,000, for which size would you order one more?

Problem D13. Do you see anything unusual in the variation illustrated in the stem and leaf plot and the relative frequency histogram? Can you think of a reason for this unusual pattern? **[See Tip D13, page 86]**



HATS SIZE (mm)

STEM	LEAF
52	2
53	1 6
54	0 5 5 5 6
55	0 1 2 5
56	0 3 4 5 8
57	3 7
58	0 0 4 5 5 6
59	0 0 0 3 4 5
60	0 0
61	3

Video Segment (approximate times: 13:31-14:45): You can find this segment on the session video approximately 13 minutes and 31 seconds after the Annenberg/CPB logo. Use the video image to locate where to begin viewing.

In this video segment, participants use a stem and leaf plot to analyze head-circumference data collected by the class. Based on what they see, they revise their initial expectations for the distribution of hat sizes. Watch this segment after completing Problem D13.

Note: The data set used by the onscreen participants is different from the one provided above.

What might one expect the middle values to be like? What accounts for the unexpected results?

Take It Further

Problem D14. Use these same data to plan an order for two more hat styles:

a. Loose Fit: Five hat sizes; hat sizes are separated by 20 mm.

b. Exclusive Fit: 20 hat sizes; hat sizes are separated by five mm.

Summary

In this session, we examined several different ways to organize continuous data measured on a number line. It is often helpful to organize this kind of data by grouping it. The stem and leaf plot is a grouping device that is useful for moderately sized data sets.

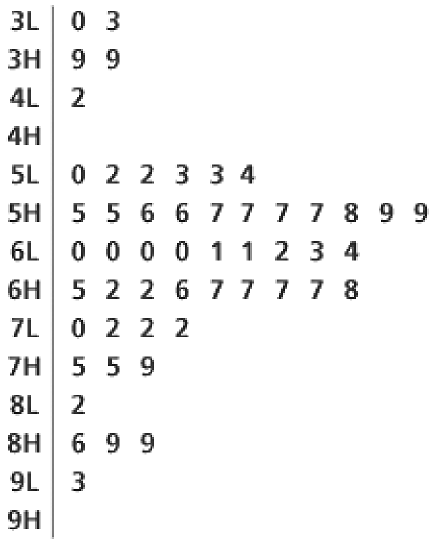
The grouped relative frequency table and histogram are more useful devices for larger data sets, since they allow you to visualize your data as portions of larger intervals. These representations, along with grouped cumulative frequency and relative frequency tables, allow you to recognize trends in large data sets by comparing the relative number of data values in each interval.

Try It Online! **www.learner.org**

Try It Online! Tabular and graphic representations of data can be reviewed online as an Interactive Activity. Go to the *Data Analysis, Statistics, and Probability* Web site at www.learner.org/learningmath and find Session 3, Part D.

Homework

This series of problems leads you through the creation of a histogram and its corresponding tables for the data in Parts B and C, which you will now group by fives. Start with the stem and leaf plot for the grouping-by-fives scenario:



3L	0 3
3H	9 9
4L	2
4H	
5L	0 2 2 3 3 4
5H	5 5 6 6 7 7 7 7 8 9 9
6L	0 0 0 0 1 1 2 3 4
6H	5 2 2 6 7 7 7 7 8
7L	0 2 2 2
7H	5 5 9
8L	2
8H	6 9 9
9L	3
9H	

Problem H1. Create a grouped frequency table for this data set where the intervals have a width of five seconds. **[See Tip H1, page 86]**

Problem H2. Create a histogram for this data set where the intervals have a width of five seconds. **[See Tip H2, page 86]**

Problem H3. Using either the histogram or the grouped frequency table, create a relative frequency table and relative frequency histogram for this data set. **[See Tip H3, page 86]**

Problem H4. Use the information from Problems H1-H3 to create a cumulative frequency and relative cumulative frequency table for this data set.

Problem H5. Using only the histogram and grouped relative frequency table based on an interval width of five, give two descriptive statements that provide an answer to the question "How well do people judge when a minute has elapsed?"

Problem H6. Based on the information in these problems, is it now possible to go back and answer any of the questions in Problem C1 that previously could not be answered with a histogram? Can you give more accurate answers for some of the questions in Problem C1? Are there some questions that still cannot be answered with a histogram?

Suggested Readings

These readings are available as downloadable PDF files on the *Data Analysis, Statistics, and Probability* Web site. Go to:

www.learner.org/learningmath

Kader, Gary and Perry, Mike (September-October, 1994). "Learning Statistics With Technology," *Mathematics Teaching in the Middle School, 1* (2), 130-136.

Pereira-Mendoza, Lionel and Dunkels, Andrejs (Summer, 1989). "Stem-and-Leaf Plots in the Primary Grades," *Teaching Statistics, 11* (2), 34-37.

Tips

Part A: Organizing Data in a Stem and Leaf Plot

Tip A6. The mean can be found by adding all the data values and dividing by the total number of values in the set.

Part B: Histograms

Tip B3. The second data set comes from a group of 52 time estimates. How many were in the first group?

Part C: Relative and Cumulative Frequencies

Tip C1. To determine whether a question can be answered, decide whether you have all the information you would need to answer it.

Tip C2. One assumption you might make is that each interval is divided evenly. So if the interval states that 15.4% of the estimates are between 80 and 90, you might assume that half of these (7.7%) are between 80 and 84 and half are between 85 and 90.

Part D: Ordering Hats

Tip D11. You will need to convert the relative frequencies into quantities of hats, adding up to 1,000. If you listed the frequencies as percentages, your data are already represented as portions of 100. Think about how you might convert your data so that they represent portions of 1,000.

Tip D13. It may help to recall that this is a unisex "Standard Fit" hat size.

Homework

Tip H1. The first interval will be 30 to < 35, the next will be 35 to < 40, etc. The last interval will be 90 to < 95.

Tip H2. If you have difficulty, refer to the guide in Part B.

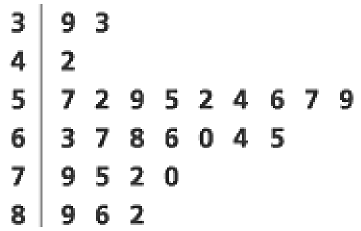
Tip H3. Refer to the guide in Part C if you have trouble here.

Solutions

Part A: Organizing Data in a Stem and Leaf Plot

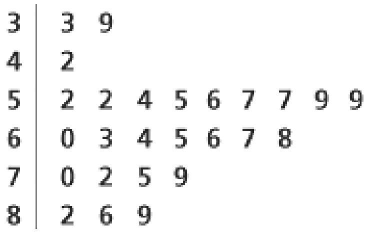
Problem A1. The range of data values is from 33 to 89, which is probably too wide for a line plot to be useful. Furthermore, there is so much variation to the data that a line plot probably would not indicate any clear trends.

Problem A2. The initial construction of the stem and leaf plot looks like this:



3	9 3
4	2
5	7 2 9 5 2 4 6 7 9
6	3 7 8 6 0 4 5
7	9 5 2 0
8	9 6 2

Problem A3. The ordered stem and leaf plot looks like this:



3	3 9
4	2
5	2 2 4 5 6 7 7 9 9
6	0 3 4 5 6 7 8
7	0 2 5 9
8	2 6 9

Problem A4.

a. All 26 estimates are between 33 and 89 seconds, as the ordered stem and leaf plot indicates.

b. Sixteen of the 26 estimates are between 52 and 68 seconds. Nine are between 52 and 59 seconds, and seven are between 60 and 68 seconds.

Problem A5.

a. There are six estimates below 55 seconds and 10 estimates above 65 seconds. In total, 16 of the 26 estimates were more than five seconds away from one minute.

b. Since 16 estimates were more than five seconds away from one minute, the remaining 10 of 26 estimates were within five seconds of one minute.

c. Three estimates were below 50 seconds, and six were above 70 seconds. In total, nine of 26 estimates were more than 10 seconds away from one minute.

d. Since nine estimates were more than 10 seconds away from one minute, the remaining 17 of 26 estimates were within 10 seconds of one minute.

Solutions, cont'd.

Problem A6.

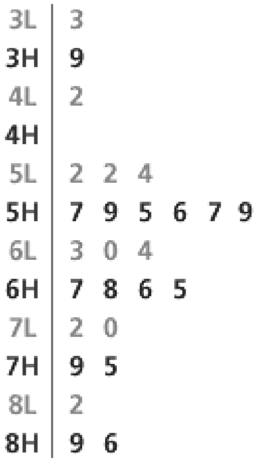
a. The median is 61.5, which is the average of the 13th data value (60) and the 14th data value (63). The mean is about 62.35, which is found by adding up all the data values and dividing by 26, the number of values in the set. (1,621 / 26 = 62.35). Both the median and mean are fairly close to 60 seconds, although we might predict that most people tend to overestimate when 60 seconds have elapsed, based on these 26 observations.

b. As the mean is 62.35, there are eight estimates above 67.35 and 10 estimates below 57.35. In total, 18 of 26 estimates are more than five seconds away from the mean.

c. There are five estimates above 72.35 and five estimates below 52.35. In total, 10 of 26 estimates are more than 10 seconds away from the mean.

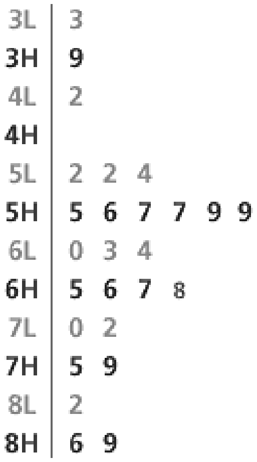
d. There is so much variability that the mode does not carry much information. In reality, it is extremely unlikely for two people to come up with *exactly* identical times for their estimates, since time is a continuous variable.

Problem A7. Before ordering, the stem and leaf plot based on a grouping by fives looks like this:



3L	3
3H	9
4L	2
4H	
5L	2 2 4
5H	7 9 5 6 7 9
6L	3 0 4
6H	7 8 6 5
7L	2 0
7H	9 5
8L	2
8H	9 6

After ordering, the stem and leaf plot based on a grouping by fives looks like this:



3L	3
3H	9
4L	2
4H	
5L	2 2 4
5H	5 6 7 7 9 9
6L	0 3 4
6H	5 6 7 8
7L	0 2
7H	5 9
8L	2
8H	6 9

Problem A8. There are many descriptive statements that could provide an answer to this question. Here are some things you may have noted in your descriptions:

- All estimates are between 33 seconds and 89 seconds. The range is 56 seconds, which indicates a lot of variation in the estimates.
- There is a concentration of estimates between 52 seconds and 68 seconds. Sixteen of the 26 estimates (or 16 / 26 = 61.5%) fall within this interval. Note that the range of this interval is only 16 seconds.
- Only two different values (57 and 59) occur more than once. There is no one value that occurs most frequently.
- The most common response time is in the range of 55 to 59 seconds, a range that contains six (23%) of the 26 estimates.
- Poor estimates are more likely to be overestimates than underestimates. Only three estimates were 50 or below, while seven estimates were 70 or higher.

Solutions, cont'd.

Problem A9.

a. A stem and leaf plot of the salaries of people working at a company or the populations of towns in a state would need wider groupings.

b. Stem and leaf plots cannot be used for qualitative data, such as gender or car color. A stem and leaf plot may not be very effective when there are few data points or when the data values are close together.

Part B: Histograms

Problem B1. A histogram offers a better graphical perspective on an entire data set. One disadvantage is that the actual data values cannot be determined from a histogram, only the number of values within intervals.

Problem B2. There are many descriptive statements that could provide an answer to this question. Here are some things you may have noted:

- All estimates are between 30 seconds and 100 seconds. The range is 70 seconds, which indicates a lot of variation in the estimates.
- There is a concentration of estimates between 50 seconds and 70 seconds. Thirty-five of the 52 estimates (or 35 / 52 = 67.3%) fall within this interval. The range of this interval is only 20 seconds.
- You may have noticed that because the histogram does not indicate individual pieces of data, we cannot look for a single number that represents the data.

Problem B3.

a. There are five estimates below 50 seconds and 12 estimates of 70 seconds or higher. In total, 17 of 52 estimates were outside the interval from 50 to less than 70 seconds.

b. Since 17 estimates were outside this interval, the remaining 35 of 52 estimates were within the interval.

c. There are four estimates below 40 seconds and five estimates of 80 seconds or higher. In total, nine of 52 estimates were outside this interval.

d. No, the answer to question (a) suggests that this group was roughly in line with the original group, since there were only 26 responses in the original group. The proportion for this group, 17 / 52 = 32.7%, is only slightly better than the proportion for the original group, which was 9 / 26 = 34.6%. Effective comparisons between groups of different sizes must be relative comparisons.

Solutions, cont'd.

Part C: Relative and Cumulative Frequencies

Problem C1.

a. Adding the percentages shows us that 90.4% of the responses are in the 70s and below.

b. Adding the percentages shows us that 9.6% are 80 or higher.

c. Adding the percentages shows us that 42.3% are in the 50s and below.

d. Adding the percentages shows us that 57.7% are 60 or higher.

e. All of them (100%) are less than 100.

f. Adding the percentages shows us that 69.2% are at least 40 but below 70.

g. This question cannot be answered using only this relative frequency table, since we are not told how many responses are in the interval 65 to < 70; we only know how many are in the interval 60 to < 70.

h. This question cannot be answered using only this relative frequency table, since we do not know how many responses are in the interval 30 to < 35.

i. This question cannot be answered using only this relative frequency table, since we do not know how many responses are exactly 60—only that 34.6% are in the interval 60 to < 70.

Problem C2. For questions (g) and (h), we might guess that half the responses will be in the lower range (60-64 and 30-34) and the other half will be in the upper range (65-69 and 35-39).

For question (g), we know that 23.1% of the responses are 70 or above and 34.6% are in the interval 60 to < 70. Half of 34.6% is 17.3%, which gives us an estimate of 23.1% + 17.3% = 40.4% for the percentage of responses that are 65 or greater.

For question (h), we know that 7.7% of the responses are in the interval 30 to < 40. Half of this is 3.85% (or 3.9%), which is our estimate for the percentage of responses less than 35.

Problem C3. Here is the completed table:

Interval	**Frequency**	**Cumulative Frequency**
30 to < 40	4	4
40 to < 50	1	5
50 to < 60	17	22
60 to < 70	18	40
70 to < 80	7	47
80 to < 90	4	51
90 to < 100	1	52

Solutions, cont'd.

Problem C4.

a. Forty-seven responses are in the 70s and below.

b. Five responses are 80 or higher (52 - 47).

c. Twenty-two responses are in the 50s and below.

d. Thirty responses are 60 or higher (52 - 22).

e. All 52 responses are less than 100.

f. Thirty-six responses are at least 40 but less than 70 (40 - 4).

Questions (g), (h), and (i) cannot be answered because the table only gives answers in intervals of 10, and these questions ask about smaller intervals.

Problem C5.

a. Using only the relative cumulative frequencies, 90.4% of the responses are in the 70s and below.

b. Using only the relative cumulative frequencies, 9.6% of the responses are 80 or higher (100 - 90.4).

c. Using only the relative cumulative frequencies, 42.3% of the responses are in the 50s and below.

d. Using only the relative cumulative frequencies, 57.7% of the responses are 60 or higher (100 - 42.3).

e. All of the responses (100%) are less than 100.

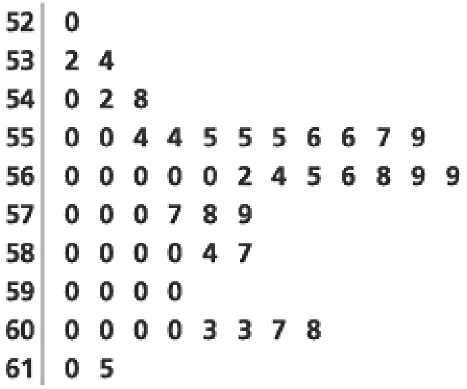
f. Using only the relative cumulative frequencies, 69.2% of the responses are at least 40 but less than 70 (76.9 - 7.7).

Part D: Ordering Hats

Problem D1. Answers will vary. One possible answer is that fewer people will have especially large or especially small head sizes, just as there are fewer people who are especially tall or short. This might suggest that the fifth and sixth (i.e., the middle) sizes would be the most common.

Problem D2. Measure your head and find out!

Problem D3. Here is the completed stem and leaf plot:



52	0
53	2 4
54	0 2 8
55	0 0 4 4 5 5 5 6 6 7 9
56	0 0 0 0 0 2 4 5 6 8 9 9
57	0 0 0 7 8 9
58	0 0 0 0 4 7
59	0 0 0 0
60	0 0 0 0 3 3 7 8
61	0 5

Solutions, cont'd.

Problem D4. Answers will vary, but here are some observations:

- All heads are between 520 and 615 mm.
- There is a range of 95 mm, which indicates a lot of variation in head circumferences.
- Thirty-five of the 55 head circumferences (63.6%) are between 550 and 587 mm, a range of 37 mm.
- Twenty-three of the 55 head circumferences (41.8%) are between 550 and 569 mm, a range of only 19 mm.

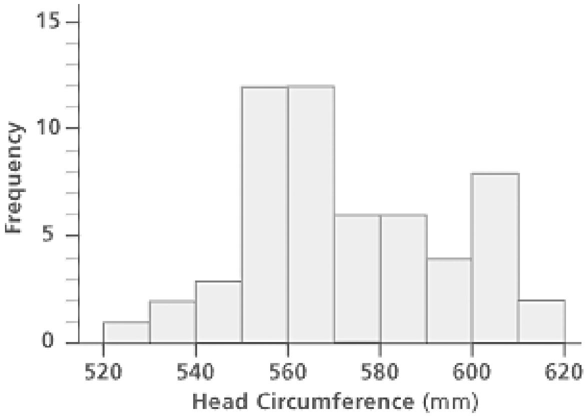
Problem D5. Head sizes between 550 and 569 mm are the most common. Head sizes below 540 mm and above 610 mm are the least common.

Problem D6. You would quickly sell out of the more common sizes and have many of the least common sizes still on hand.

Problem D7.

Head Circumference (mm)	Frequency	Relative Frequency (%)
520 - < 530	1	1.8
530 - < 540	2	3.6
540 - < 550	3	5.5
550 - < 560	11	20.0
560 - < 570	12	21.8
570 - < 580	6	10.9
580 - < 590	6	10.9
590 - < 600	4	7.3
600 - < 610	8	14.5
610 - < 620	2	3.6

Note that the relative frequencies add up to 99.9%, due to rounding.



Solutions, cont'd.

Problem D8. You no longer have the actual data values, only the number of values within intervals of 10 millimeters.

Problem D9. Answers will vary, but here are some observations:

- All heads are between 520 and 620 mm.
- There is a range of 100 mm, which indicates a lot of variation in head circumferences.
- Thirty-five of the 55 head circumferences (63.6%) are between 550 and 590 mm, a range of 40 mm.
- Twenty-three of the 55 head circumferences (41.8%) are between 550 and 570 mm, a range of only 20 mm.

Problem D10. Head sizes between 550 and 570 mm are the most common. Head sizes below 540 mm and above 610 mm are the least common.

Problem D11. Perform this by expressing the relative frequency as a decimal, then multiplying this decimal by 1,000. (If you wanted to work with the percentage value without converting it to a decimal, you need to remember that percentages are per 100, so you would need to multiply the percentage value by 10 to find the number per 1,000.)

Size	Number To Order
S1	18
S2	36
S3	55
S4	200
S5	218
S6	109
S7	109
S8	73
S9	145
S10	36

Problem D12.

a. You should have found a total of only 999 hats, due to the rounding in the relative frequencies from Problem D7.

b. Answers will vary. One possible answer is to use S4 or S5, since they are the most common sizes. Another is to use either S3 or S9, since the numbers of hats in these sizes when written as decimals are closest to being rounded up (S9, for example, would be 145.4545... hats).

Problem D13. Yes. There are two distinct peaks in the histogram**,** which may be due to the fact that male and female head sizes are mixed together in this data set. This raises several questions: Do men and women have similar-sized heads? If not, do men tend to have larger heads than women, or do women tend to have larger heads than men?

Problem D14. To calculate these answers, you will first need to set the hat sizes, then use the data values to determine the relative frequency of the hat sizes you selected, then multiply these frequencies expressed as decimals by 1,000 to determine how many of each you will order. Answers will vary, due to the flexibility in selecting the intervals for the hat sizes.

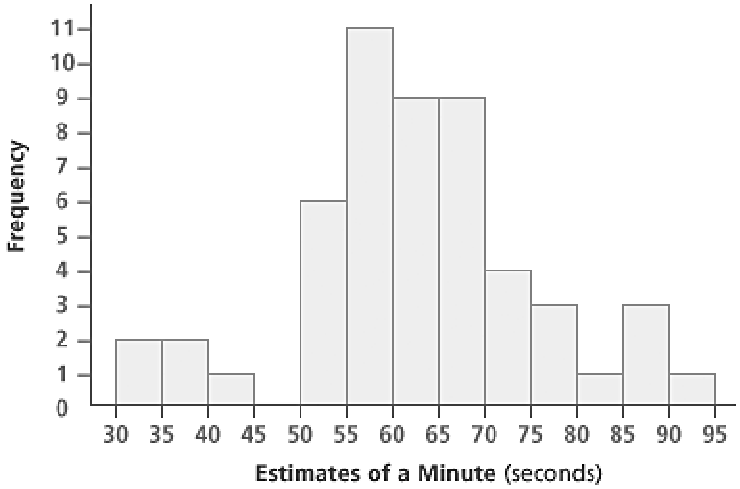
Solutions, cont'd.

Homework

Problem H1.

Interval	Frequency for Interval
30 to < 35	2
35 to < 40	2
40 to < 45	1
45 to < 50	0
50 to < 55	6
55 to < 60	11
60 to < 65	9
65 to < 70	9
70 to < 75	4
75 to < 80	3
80 to < 85	1
85 to < 90	3
90 to < 95	1

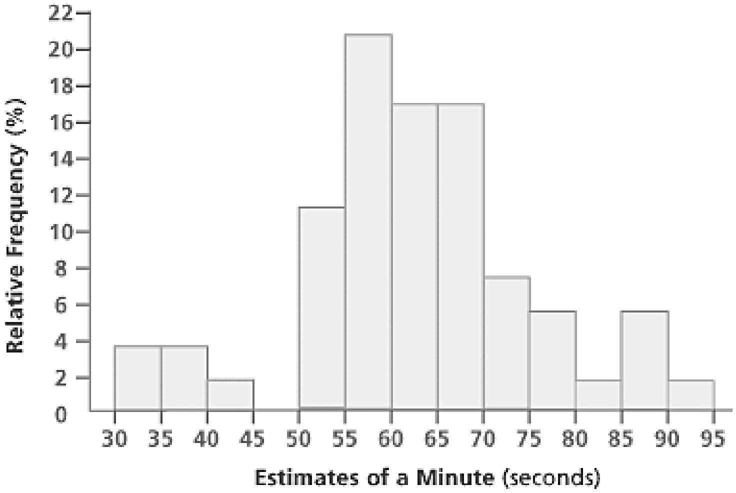
Problem H2.



Solutions, cont'd.

Problem H3.

	Relative Frequency		
Interval	**Fraction**	**Decimal**	**Percentage**
30 to < 35	2 / 52	.038	3.8
35 to < 40	2 / 52	.038	3.8
40 to < 45	1 / 52	.019	1.9
45 to < 50	0 / 52	.000	0.0
50 to < 55	6 / 52	.115	11.5
55 to < 60	11 / 52	.212	21.2
60 to < 65	9 / 52	.173	17.3
65 to < 70	9 / 52	.173	17.3
70 to < 75	4 / 52	.077	7.7
75 to < 80	3 / 52	.058	5.8
80 to < 85	1 / 52	.019	1.9
85 to < 90	3 / 52	.058	5.8
90 to < 95	1 / 52	.019	1.9



Solutions, cont'd.

Problem H4.

		Relative Cumulative Frequency	
Interval	**Cumulative Frequency**	**Fraction**	**Percentage**
30 to < 35	2	2 / 52	3.8
35 to < 40	4	4 / 52	7.7
40 to < 45	5	5 / 52	9.6
45 to < 50	5	5 / 52	9.6
50 to < 55	11	11 / 52	21.2
55 to < 60	22	22 / 52	42.3
60 to < 65	31	31 / 52	59.6
65 to < 70	40	40 / 52	76.9
70 to < 75	44	44 / 52	84.6
75 to < 80	47	47 / 52	90.4
80 to < 85	48	48 / 52	92.3
85 to < 90	51	51 / 52	98.1
90 to < 95	52	52 / 52	100.0

Problem H5. All of the response times are between 30 and < 95 seconds. There is a concentration of responses (42 of 52) between 50 and < 80 seconds. The most frequently occurring interval is from 55 to < 60 seconds, followed by both 60 to < 65 seconds and 65 to < 70 seconds.

Problem H6. Yes, questions (g) and (h) can now be answered, since the intervals are broken up by fives. The only question that still cannot be answered is (i), the percentage of responses that are equal to 60 seconds.